行为可见增加利他偏好及其社会规范机制[1]

黄馨茹[1]　李　健[1,2]　倪荫梅[1]

（[1]心理与认知科学学院，行为与心理健康重点实验室，北京大学，北京，100871）

（[2]IDG 麦戈文脑科学研究所，北京大学，北京 100871）

摘要

在不同社会偏好类型中，研究者较为关注利他偏好及其信号功能。本研究探究在独裁者游戏中，决策者的利他偏好如何受到分配方案对接受者可见性的影响。实验一采用行为实验结合计算建模的方法，发现无论在选择或评分条件下，相比于行为不可见，当分配者的行为能够为接受者所见时，分配者都表现出更大程度的利他偏好。此外，相比于评分条件，在选择时人们更加在意分配效率。实验二仅使用选择条件，并操纵社会规范，发现行为可见增加利他偏好的作用依赖于利他的社会规范，当存在非利他社会规范时，行为可见的影响减小。本研究结果表明，在利他社会规范下，当行为对接受者可见时，人们将表现出更多利他偏好。

关键词

社会偏好，利他，行为可见性，反应类型，社会规范

1 前言

在涉及人际互动的经济决策中，个体的行为往往不仅取决于自己的潜在收益，还受他人可能收益的影响。社会偏好(social preference)指相比于自我收益，考虑他人收益的程度和方向(Glimcher & Fehr, 2013)。社会偏好包括多种类型，如社会价值取向(social value orientation)理论将社会偏好分为利他、亲社会、个人主义和竞争四类，其中亲社会又可分为追求总收益最大化和追求公平(Murphy et al., 2011; van Lange, 1999; van Lange et al., 1997)。在不同社会

---

[1] 收稿日期：2021-12-20
国家自然科学基金委员会面上基金（31871140，32071090），国家科技创新 2030 重大项目（2021ZD0203700 / 2021ZD0203704），中国博士后科学基金（2021M690238），倪荫梅受北京大学-清华大学生命科学联合中心博士后基金资助。
共同通讯作者：李健，E-mail: li.jian@pku.edu.cn; 倪荫梅，E-mail: niyinmei@pku.edu.cn

偏好类型中，利他偏好(altruistic preference)受到研究者较多关注。社会价值取向理论将利他偏好定义为“不在意自我收益，只希望最大化他人收益”(Murphy et al., 2011)。然而在较多情况下，利他者并非完全不考虑自我收益(Pfattheicher et al., 2022; West et al., 2007)。因此，利他偏好更多地被定义为“对他人收益有正向的考虑”(Sáez et a., 2015)，即，在其他因素保持不变的情况下，只要个体希望增加他人的收益，便认为该个体表现出利他偏好。

较多前人研究对利他的原因与机制进行探讨并提出理论解释。其中，一系列理论指出利他具有信号功能。如，竞争利他假说(competitive altruism hypothesis, Hardy & Van Vugt, 2006)和高成本信号理论(costly signaling theory, Gintis et al, 2001)认为人们竞争性地进行利他行为，向群体传递个体亲社会的信号并建立较好的社会形象，进而获得声誉和地位并在其后的社会互动中获益。该过程中，利他行为能够被他人看见是利他发挥信号功能的关键步骤。本研究将从利他的信号功能出发，探究行为可见性对于利他行为的影响。

一些研究发现，在各类社会互动中，人们的行为较容易受到他人是否知晓或观察其行为的影响(e.g., Dana et al., 2006; Fox & Guyer, 1978; Hamilton & Lind, 2016; Jerdee & Rosen, 1974; Lacetera & Macis, 2010)。Bradley等人(2018)将“可观察性(observability)”分为程度由浅至深的三类：第一类为“感知到被观察”，即仅采用视觉线索对被试施加被观察感，但并不存在真实的观察者(e.g., Bateson et al., 2006; Burnham & Hare, 2007; Haley & Fessler, 2005; Oda et al., 2011; Raihani & Bshary, 2012)；第二类为“行为被观察”，即他人可以知晓该行为的发生，但并不知晓行为者的身份(即不知晓行为者的姓名等身份信息) (e.g., Andreoni & Bernheim, 2009)；最后一类为“行为和身份均被观察”(e.g., Andreoni & Petrie, 2004; McAuliffe et al., 2013; 2020)。已有的大部分研究集中于第一、三类可观察性，而对生态效应较广的第二类研究尚较缺乏。

另一方面，有研究表明观察者的不同身份，如作为利他行为的接受方(利益攸关方)或旁观的第三方也可能对行为可见性效应造成影响(McAuliffe et al., 2020)。目前较多研究关注第三方对个体利他行为的影响，如“第三方惩罚”相关研究(Fehr & Fischbacher, 2004; McAuliffe et al., 2015)发现，当分配者知晓第三方存在时，会倾向于做出更加利他的决策(田莹，2016)；而较少讨论行为对接受方是否可见对于利他行为的影响(Andreoni & Bernheim, 2009)。相比于第三方，接受方的利益直接受到分配者行为的影响，因此，相比于第三方可见，当行为对接受者可见时，分配者可能表现出不同的利他行为变化。在少量此类研究如Andreoni和Bernheim(2009)的实验中，分配者知道每一试次是否由自己决定，但接受者只知道分配由分配者决定的概率。当分配100%由分配者做出时，分配者将会选择他人收益更多的选项；而

当分配结果只有较小概率取决于分配者行为时，分配者的选择将会更加利己。这项研究说明接受者对行为可知能够增加分配者的利他行为，然而这项研究以概率操纵可见性，较不直观，与现实社会的情境相差较大。综上所述，区别于已有大量考察被观察感、行为和身份均被观察或利益无关第三方观察者的研究范式，本研究关注当观察者为接受方时，行为是否可见对于分配者利他行为的影响。

为了对社会偏好进行量化，本研究在社会决策的研究框架下，从计算加工机制的角度，对决策的效用函数进行计算建模。Fehr和Schmidt(1999)提出的“不公平厌恶(inequity aversion)”模型，可以较好地对社会偏好进行量化研究，并揭示社会决策的认知加工机制。其效用函数表达式如下：

$$U = M_s - \alpha \cdot max\{M_s - M_o, 0\} - \beta \cdot max\{M_o - M_s, 0\} \quad (1)$$

该模型认为，个体在决策时不仅考虑自我的净收益(M_s)，还考虑分配的公平程度($M_s - M_o$)，即自我他人收益之差；人们既希望最大化自我收益，又希望最小化自我他人收益之差。具体而言，当自我收益多于他人收益时，个体表现出优势不公平厌恶(advantageous inequity aversion, AIA，α)；当自我收益少于他人收益时，个体表现出劣势不公平厌恶(disadvantageous inequity aversion, DIA，β)。Fehr-Schmidt模型认为人们在优势和劣势不公平情境下都表现出追求公平的偏好，即AIA和DIA均大于零，这一结果得到部分研究的支持(e.g., Hu et al., 2018; Sáez et al., 2015)。然而，也有研究汇报DIA与AIA小于零的情形(Gao et al., 2018)。Sáez等人(2015)认为，当AIA大于零且DIA小于零时，表明人们既不希望自我收益多于他人，也接受他人收益多于自己，这时候AIA和DIA均能指示利他偏好，反映人们希望增加他人相对收益的倾向性。

在Fehr-Schmidt模型框架下，结合AIA和DIA的正负取值，可以在不同象限中表示多种社会偏好(图1)：AIA > 0且DIA > 0，对应公平偏好；AIA > 0且DIA < 0，对应利他偏好(Sáez et al., 2015)；AIA < 0且DIA > 0，说明个体能接受自己收益多于他人，且不希望他人收益多于自己，即对他人收益的相对权重为负，表现为非利他偏好；AIA < 0且DIA < 0，在特定的实验设计中(Gao et al., 2018)，AIA为负代表希望增加自我收益，DIA为负代表希望增加他人收益，总体表现为在意分配效率。本研究将沿用Gao等人(2018)的实验设计，据此定义AIA和DIA在四个象限的正负取值对应的社会偏好，并在此框架下提出研究假设一：接受者对决策者行为可见能够增加决策者的利他偏好；相比于不可见条件，在可见条件下，AIA上升，DIA下降。

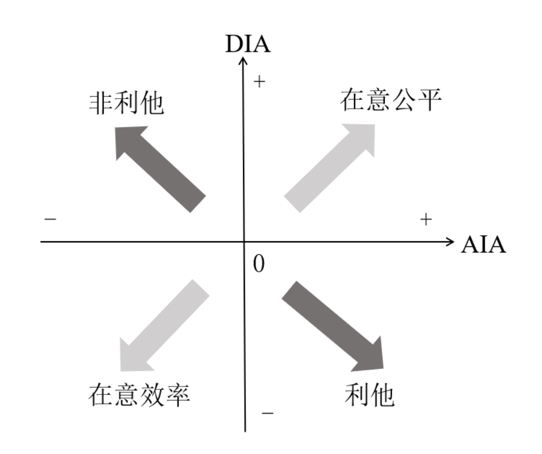


图1 AIA和DIA正负方向对应的社会偏好

为定量研究社会偏好，需对社会偏好的测量方式进行谨慎选择。效用理论认为偏好是人们对于选项的内在倾向程度，直接影响其决策行为(Raiffa & Schlaifer, 1961)。因此，通过观察人们外显的、离散的决策行为，可以反向推导和测量人们内在的、连续的偏好。关于社会偏好的测量，目前有两种主要的行为反应类型：通过选择决策反推偏好(Gao et al., 2018; Sáez et al., 2015)和通过主观评价(评分)测量偏好(Liu et al., 2019; Loewenstein et al., 1989)。实验一分别采用选择和评分两种方式测量社会偏好，探讨行为可见性对于利他偏好的影响是否具有跨反应类型的稳定性，并据此提出研究假设二：在选择和评分两种反应类型中，行为可见都会增加利他偏好。

前人关于可见性影响利他行为的机制研究尚不丰富，因此，本研究还希望探索社会规范对于行为可见增加利他偏好的作用。社会规范分为描述性规范(描述大多数人实际的行为模式)和命令性规范(指示大多数人认可的正确行为模式) (Cialdini et al., 1990)。研究发现，无论在现实情境中(e.g., Dempsey et al., 2018; Palacios et al., 2022)，还是实验室情境下(e.g., Kawamura & Kusumi, 2017; Raihani & McAuliffe, 2014)，向被试传达描述性规范都能够明显改变被试的行为，使其向社会规范趋近，故本研究主要关注描述性社会规范。

社会规范在可见的社会决策中可能具有重要作用。首先，有研究者认为社会规范极大程度驱动了亲社会偏好，并发现不同的社会规范遵从性能够解释不同的社会偏好(Kimbrough & Vostroknutov, 2016; McBride & Ridinger, 2021)。并且，一项研究发现向被试传达利他的社会规范能够显著增加被试的利他行为(Agerström et al., 2016)。其次，行为可见和社会规范也可能存在联系，如，当人们感到被观察时，其行为变异性减小(Nettle et al., 2013)，提示行为可见时人们的行为可能更倾向于遵从某种特定规范。一些其他类型的可见性研究也提到了社会规范的作用。如，Fehr和Fischbacher(2004)认为在第三方可见时，第三方的存在维护了公平

和合作的规范；在关于“感知到被观察”的研究中，发现若给予非利他的社会规范，那么被观察的视觉线索不会增加被试的捐赠数额(Kawamura & Kusumi, 2017)。然而，在本研究关注的情境中，即接受者对决策者行为可见增加决策者利他偏好的过程中，社会规范的影响未知。基于前人研究，社会互动中普遍存在利他社会规范(Pereda et al., 2017)，我们推测行为可见可能让被试更易感知这种规范，进而做出更多利他行为；那么，当利他规范不存在或消失时，行为可见对利他偏好的影响将减小。基于此，本研究将在实验二中探究并验证假设三：行为可见增加利他偏好的效果受到利他规范存在与否的影响，具体而言，相比于不存在利他规范，利他规范下行为可见增加利他偏好的程度更大。

综上所述，本研究将用两个实验验证三个假设。实验一和实验二均采用修改版独裁者游戏营造社会决策情境，并使用不公平厌恶模型量化社会偏好，探究行为可见性对社会偏好的影响，以验证假设一；此外，实验一分别采用选择(被试需要选择以决定自我和他人收益)和评分(被试需要对给定自我和他人收益的分配方案做出评价)两种方式测量社会偏好，探讨行为可见性的影响在不同反应类型间的稳定性，以验证假设二；实验二通过操纵社会规范，探究行为可见增加利他偏好的效应如何受到利他或非利他的社会规范的影响，以验证假设三。

2 实验一

2.1 方法

2.1.1 被试

收集数据前，使用 G*Power 3.1.9.7 计算预期样本量，对于实验一的设计，基于 2×2 重复测量方差分析的 F 检验，当效力为 0.8、显著性水平为 0.05、效应量为 0.25 时，所需样本量为 34。实际研究中，共招募 41 名成年大学生被试(男性 16 人)，其中 3 人因数据记录不完整而被排除，剩余有效被试 38 人(男性 14 人)，年龄在 19~30 岁之间($M = 22.03$，$SD = 2.70$)。所有被试母语均为汉语，右利手，视力或矫正视力正常。所有被试均在实验前填写知情同意书，知晓实验的潜在收益与风险，自愿参与实验，并在实验后获得金钱报酬。

2.1.2 实验设计与流程

实验一为 2(行为可见性：可见 vs. 不可见) × 2(反应类型：选择 vs. 评分) × 2(不公

平类型：优势 vs. 劣势)的被试内设计。自变量为行为可见性(简称可见性)、反应类型和不公平类型。因变量为被试的利他程度，由被试在选择条件下的选择行为和评分条件下的主观评分测量。

实验开始前，4~8 名同性别的被试同时来到实验室(4 人、6 人或 8 人)。被试被告知接下来的人际互动实验中，在场的人将有一半被随机指定成为分配者，另一半为接受者，每一轮游戏均由一名分配者和一名接受者随机配对完成(每一轮游戏开始前会重新配对)，所有游戏匿名进行。事实上，所有人都会成为分配者，实验中的“搭档”并不真实存在(实验后的真实性测试表明被试相信搭档真实存在)。被试通过电脑程序完成实验，实验程序由 MATLAB® R2018a 编写。

实验任务为区块设计，共 4 个区块，分别为可见-选择，可见-评分，不可见-选择和不可见-评分。其中，选择的两个区块和评分的两个区块总是同时出现，但选择和评分的前后顺序在被试间随机进行；且选择或评分内部，可见与不可见条件的前后顺序随机呈现。每一区块开始前，屏幕提示该区块属于可见或不可见条件以及被试进行选择或评分反应。

每一轮游戏中，被试和“搭档”首先完成一道简单的加减法计算，若两人均回答正确，则进入奖励分配，并进行 4 次选择或 4 次评分，否则二人将等待 20s 后进入下一轮。这一设置的目的在于减少初始投入与任务贡献对于分配倾向性的影响(Li et al., 2018)。

选择条件下，被试作为分配者在两个选项中选择一个，以决定自己和“搭档”这一轮获得的分数。其中，一个选项总是公平选项(自己得到 10 分，“搭档”得到 10 分)，另一个选项是不公平选项(一半为优势不公平，如“自己得到 12 分，‘搭档’得到 8 分”，另一半为劣势不公平，如“自己得到 8 分，‘搭档’得到 12 分”)。选项设置参考 Gao 等人(2018)的研究。评分条件下，屏幕将呈现一个公平和一个不公平方案，并由程序选中一个，该方案决定了被试和“搭档”该轮分数。被试被告知他们和“搭档”需要各自对这个分配的满意程度在-4 ~ 4 的 9 点量表上进行评分。

被试做出反应后，进入反馈阶段。可见条件下，被试被告知他们的选择/评分结果正在展示给搭档，由睁开的眼睛表示；不可见条件下，“搭档”不会看到这一轮被试的选择/评分结果，由闭合的眼睛表示。每一试次的实验流程如图 2 所示。

选择区块共 12 轮游戏，共 12×4=48 次选择。评分区块共 18 轮游戏，共 18×4=72 次评分，其中，48 次程序选择不公平分配选项(和选择条件一一对应)；为使实验设计显得合理，还有 24 次程序选择公平分配的选项。

最后，被试在实验中获得的所有分数将会累加并以 60:1 的比例转换为实验报酬（每个

被试获得约 60 元的报酬）。

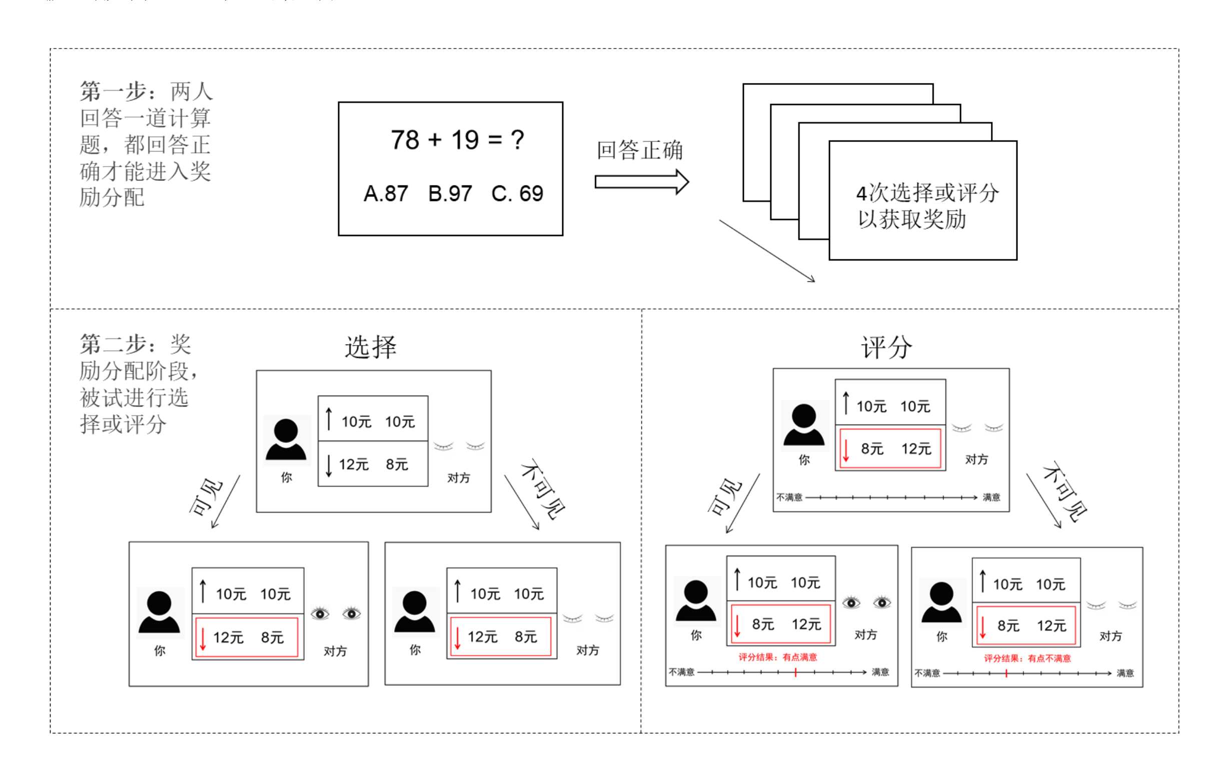


图 2 实验一流程

2.1.3 测量指标

实验因变量为利他程度，通过不依赖模型的指标测量外显的利他选择/评分，并使用计算建模揭示被试内隐的利他偏好。

在选择条件下，参照前人研究(Gao et al., 2018; Sáez et al.,2015)，计算每名被试选中选项的自我与他人收益差异的绝对值均值(简称自我他人收益之差)：若在优势不公平条件，自我他人收益之差越小，在劣势不公平条件，自我他人收益之差越大，说明被试越不希望自己收益多于他人，同时越能容忍他人收益多于自己，表现出更高的利他程度。在评分条件下，参照前人研究(Liu et al., 2019)，使用对不公平选项的评分均值作为行为指标：若对优势不公平选项评分越低，对劣势不公平选项评分越高，则说明更加偏好优势公平和劣势不公平，表现出更高的利他程度。

计算建模部分使用不公平厌恶模型(Fehr & Schmidt, 1999)对被试的偏好进行建模。假定选择和评分条件下，被试对不公平选项的效用均为：

$$U_{unequal} = \begin{cases} M_s - \alpha_1 \cdot max(M_s - M_o,\ 0) - \beta_1 \cdot max(M_o - M_s, 0) & \text{可见} \\ M_s - \alpha_2 \cdot max(M_s - M_o,\ 0) - \beta_2 \cdot max(M_o - M_s, 0) & \text{不可见} \end{cases} \quad (2)$$

该模型假设被试的效用由自我收益M_s和自我他人收益之差$(M_s - M_o)$两部分组成。其中，

α_1和β_1分别代表可见条件下的 AIA 和 DIA，α_2和β_2分别代表不可见条件下的 AIA 和 DIA。AIA 越大，DIA 越小，说明利他程度越高。

对于公平选项(M_s = 10，且M_o = 10)，被试的效用为：

$$U_{equal} = 10 \quad (3)$$

在选择和评分条件下，主观效用以不同方式体现在被试的外显行为上。对于选择条件，假设被试对不公平选项的选择服从参数为$P_{unequal}$的伯努利分布，其中$P_{unequal}$等于两个选项主观效用之差的 SoftMax 转换：

$$P_{unequal} = \frac{1}{1+e^{-\lambda(U_{unequal}-U_{equal})}} \quad (4)$$

等式中，λ为温度参数，反映被试的选择对决策效用(不公平与公平选项效用差)的敏感性。假设可见与不可见条件下的λ不同，记为λ_1、λ_2。

对于评分条件，假设评分的分值R和两个选项的主观效用之差存在线性关系(Liu et al., 2019)，被试对不公平选项和公平选项的评分分别满足：

$$R_{unequal} = b_0 + b_1 \cdot (U_{unequal} - U_{equal}) + \varepsilon \quad (5)$$

$$R_{equal} = b_0 + b_1 \cdot (U_{equal} - U_{unequal}) + \varepsilon \quad (6)$$

其中，b_0为截距、b_1为斜率，ε为误差项(ε~$N(0,e)$)。假设可见与不可见条件下的b_0、b_1与e不同，分别记为b_{01}、b_{11}、e_1和b_{02}、b_{12}、e_2。

使用 R 4.0.2 与 JAGS 4.3.0 建立分层贝叶斯模型(Hierarchical Bayesian Model, HBM)进行参数拟合。假定单个被试在可见与不可见条件下的参数分布存在重合，即$\alpha_2 = \alpha_1 + \triangle\alpha$，$\beta_2 = \beta_1 + \triangle\beta$，$\lambda_2 = \lambda_1 + \triangle\lambda$，$b_{02} = b_{01} + \triangle b_0$，$b_{12} = b_{11} + \triangle b_1$。在选择条件下，对于每名被试，同时估计 6 个参数：α_1、$\triangle\alpha$、β_1、$\triangle\beta$、λ_1、$\triangle\lambda$。在评分条件下，对于每名被试，同时估计 10 个参数：α_1、$\triangle\alpha$、β_1、$\triangle\beta$、b_{01}、$\triangle b_0$、b_{11}、$\triangle b_1$、e_1、e_2。HBM 模型中，个体参数来自于群体正态分布，对于每个参数，其群体正态分布的均值的先验分布均为参数为(0, 1)的正态分布，精度(方差的倒数)的先验分布均为参数为(0.1, 0.1)的伽马分布。得到参数的后验分布后，取后验分布的中位数作为参数估计值。共建立两个模型，并进行模型比较：在选择-评分联合模型中，选择和评分采用相同的α与β参数值；在选择-评分分离模型中，选择和评分采用不同的α与β参数值。

在建模之前，我们还验证了在评分条件下，两个选项都对被试的评分产生影响。使用混合线性模型，以评分值为因变量，以两个选项中的自我和他人收益(分别记为Ms_1、Mo_1、Ms_2、Mo_2)以及可见性($visibility$)为固定效应，并加入个体随机效应截距，建立如下回归模型：

$$Rating = \beta_{0i} + \beta_1 Ms_1 + \beta_2 Mo_1 + \beta_3 Ms_2 + \beta_4 Mo_2 + \beta_5 visibility + u \quad (7)$$

其中，u代表服从标准正态分布的误差项，β_{0i}表示随机截距，β_1~β_5分别代表不同自变量的效应。

2.2 结果

本研究使用 IBM SPSS Statistics 20、MATLAB® R2018a 和 R 4.0.2 进行数据分析与统计检验，结合不依赖模型的结果和计算建模以验证研究假设。

2.2.1 不依赖模型的结果

分别分析行为可见性对选择和评分条件下利他程度的影响。对于选择条件，以自我他人收益之差(绝对值)为因变量，以不公平类型(优势 vs. 劣势)和可见性(可见 vs. 不可见)为自变量，进行重复测量方差分析。结果显示，不公平类型的主效应显著，$F(1,37) = 177.59$，$p < 0.001$，$\eta_p^2 = 0.83$，相比劣势，在优势条件下自我他人收益之差更大；可见性的主效应不显著，$F(1,37) = 2.63$，$p = 0.113$，$\eta_p^2 = 0.066$；不公平类型与可见性的交互作用显著，$F(1,37) = 22.00$，$p < 0.001$，$\eta_p^2 = 0.37$。简单主效应分析发现，在优势条件下，可见条件下自我他人收益之差显著小于不可见条件($D = -1.07$，$p < 0.001$) (图 3a)；在劣势条件下，可见条件下的自我他人收益之差显著大于不可见条件($D = 0.544$，$p = 0.041$) (图 3b)。

对于评分条件，以评分值为因变量，以不公平类型(优势 vs. 劣势)和可见性(可见 vs. 不可见)为自变量，进行重复测量方差分析。结果显示，不公平类型的主效应显著，$F(1,37) = 84.04$，$p < 0.001$，$\eta_p^2 = 0.69$，相比劣势，在优势条件下评分值更大；可见性的主效应不显著 $F(1,37) = 1.66$，$p = 0.205$，$\eta_p^2 = 0.043$；不公平类型与可见性的交互作用显著 $F(1,37) = 48.81$，$p < 0.001$，$\eta_p^2 = 0.57$。简单主效应分析发现，在优势条件下，可见条件下评分值显著小于不可见条件($D = -0.43$，$p < 0.001$) (图 3c)；在劣势条件下，可见条件下的评分值显著大于不可见条件($D = 0.60$，$p < 0.001$) (图 3d)。

以上结果揭示，可见性对利他程度的影响在选择和评分间一致：相比于不可见条件，可见条件下更不能容忍优势不公平，更能容忍劣势不公平，总体表现出更高的利他程度。

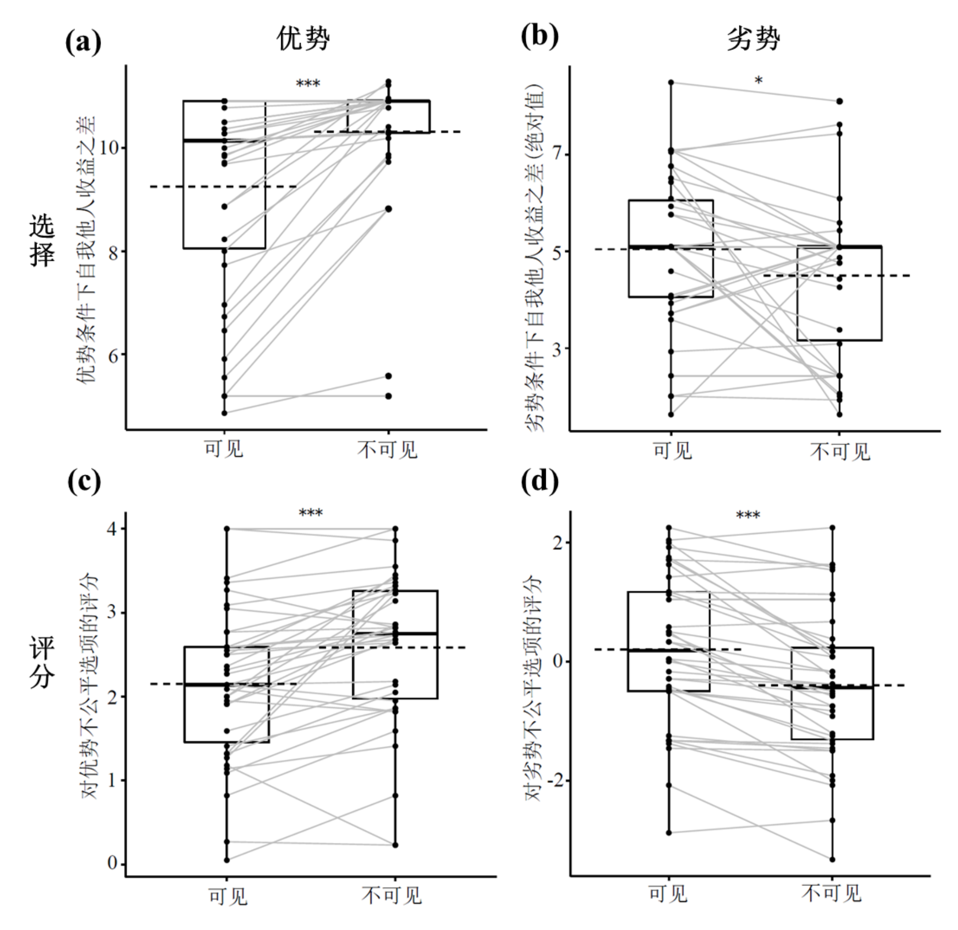


图 3 实验一不依赖模型的结果 (a)(b)选择条件，优势和劣势情况下的自我他人收益之差；**(c)(d)** 评分条件，优势和劣势情况下的评分值。箱图内部的水平实线代表中位数，水平虚线代表均值。*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$。

2.2.2 计算建模结果

在进行计算建模之前，首先验证建模方式的合理性。使用方法部分的(6)式进行回归分析发现，被试的评分同时受到已选和未选选项的自我和他人收益影响($\beta_1 = 0.18$，$t = 55.35$，$p < 0.001$；$\beta_2 = -0.016$，$t = -4.51$，$p < 0.001$，$\beta_3 = -0.10$，$t = -22.12$，$p < 0.001$；$\beta_4 = -0.030$，$t = -6.47$，$p < 0.001$)。因此，对评分条件的效用建模同时考虑已选选项和未选选项。模型比较结果显示，选择评分分离模型(DIC = 21233.43)优于选择-评分联合模型(DIC = 22185.17)。选择-评分分离模型的参数拟合结果如图 4(a)和(b)所示。

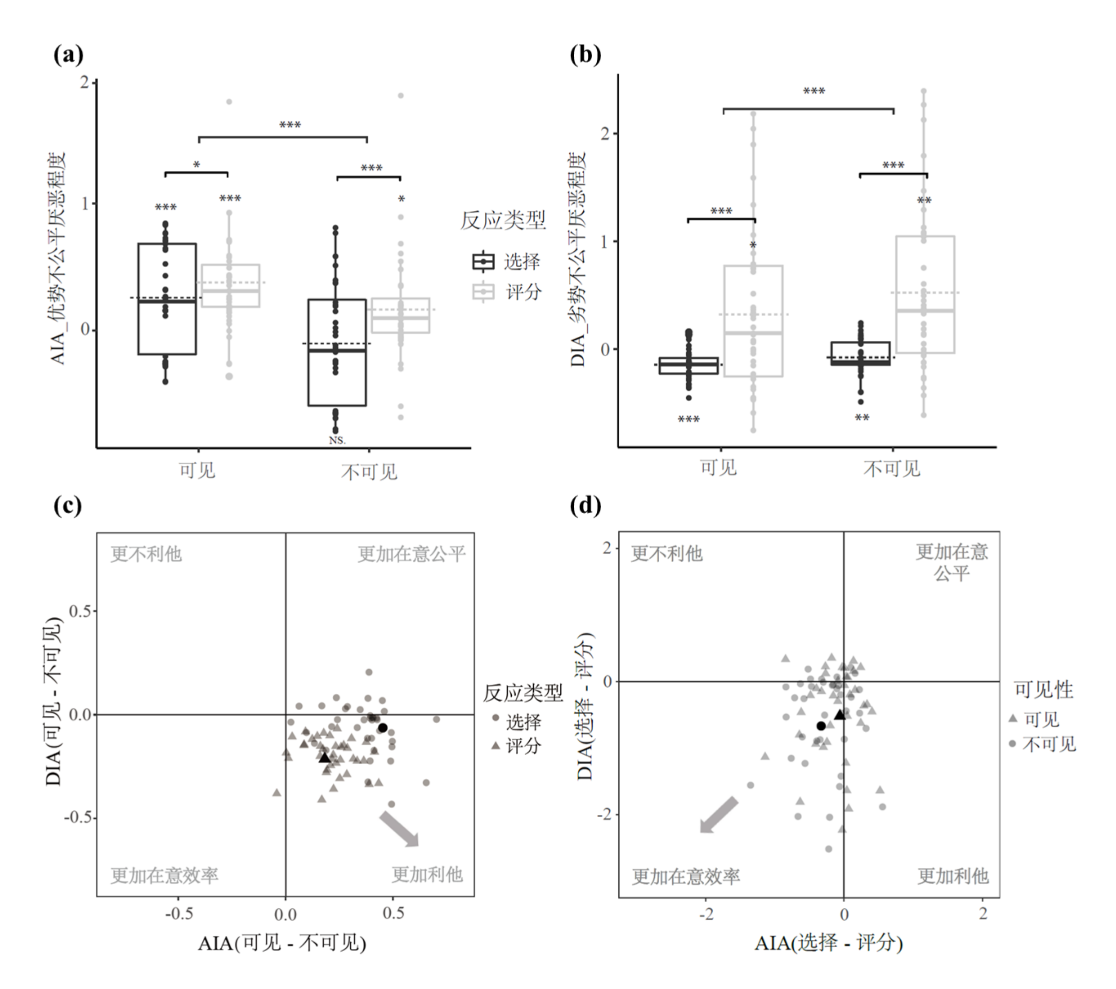


图 4 实验一基于模型的结果 (a) (b)模型估计的 AIA 和 DIA；**(c)**选择和评分条件下，可见与不可见条件之间的 AIA 之差与 DIA 之差；**(d)**可见和不可条件下，选择与评分条件之间的 AIA 之差与 DIA 之差。(a)(b)中，每个箱图上方的显著性表示该条件下参数显著大于零，下方的显著性表示该条件下参数显著小于零。图像最上方的显著性表示可见与不可见情况下，选择和评分条件的差值的比较。箱图内部的水平实线代表中位数，水平虚线代表均值。*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$。(c) (d)中，浅色图案为个体层面模型拟合结果，深色图案为各个条件下的均值。

为检验不同可见性与反应类型下 AIA 和 DIA 的差异，分别以 AIA、DIA 为因变量，以可见性和反应类型为自变量，进行 2×2 重复测量方差分析。以 AIA 为因变量进行方差分析发现，可见性的主效应显著，$F(1,37) = 255.17$，$p < 0.001$，$\eta_p^2 = 0.87$，可见条件下 AIA 大于不可见条件；反应类型的主效应显著，$F(1,37) = 11.15$，$p = 0.002$，$\eta_p^2 = 0.23$，选择条件下 AIA 小于评分条件；可见性与反应类型的交互作用显著，$F(1,37) = 38.53$，$p < 0.001$，$\eta_p^2 = 0.55$。简单主效应分析发现，在不可见条件下，选择条件下的 AIA 值显著小于评分条件($D = -0.26$，$p < 0.001$)；在可见条件下，选择条件下的 AIA 值也略小于评分条件($D = -0.13$，$p = 0.032$)。

以 DIA 为因变量进行方差分析发现，可见性的主效应显著，$F(1,37) = 123.49$，$p < 0.001$，$\eta_p^2 = 0.77$，可见条件下 DIA 小于不可见条件；反应类型的主效应显著，$F(1,37) = 22.01$，$p < 0.001$，$\eta_p^2 = 0.37$，选择条件下 DIA 小于评分条件；可见性与反应类型的交互作用显著，$F(1,37) = 32.13$，$p < 0.001$，$\eta_p^2 = 0.47$。简单主效应分析发现，在不可见条件下，选择条件下的 DIA 值显著小于评分条件($D = -0.60$，$p < 0.001$)；在可见条件下，选择条件下的 DIA 值也显著小于评分条件($D = -0.46$，$p < 0.001$)。

图 4(c-d)分别直观地描述了可见性和反应类型对 AIA 和 DIA 的影响。相比于不可见条件，在可见条件下 AIA 增大，DIA 减小，总体向第四象限偏移，表明行为可见增进利他偏好(图 4c)；相比于评分条件，在选择条件下 AIA 减小，DIA 也减小，总体向第三象限偏移，表明选择时被试更在意分配效率(图 4d)。

计算建模结果中可见性与反应类型的交互作用提示，可见性缩小选择和评分条件下的参数差异。为了进一步验证该结果，分别计算选择条件和评分条件下 AIA 的差值和 DIA 的差值，并对该差值在可见与不可见条件下进行配对样本 t 检验。结果发现，相比于不可见条件，在可见条件下，选择与评分条件间的 AIA 差值($t = -6.21$，$p < 0.001$，Cohen'$d = 0.37$)和 DIA 差值($t = -5.67$，$p < 0.001$，Cohen'$d = 0.20$)都更小。该结果说明，相对于行为不可见，行为可见能够缩小同一个体选择和评分条件下社会偏好的差异性，使个体的社会偏好更趋一致。

3 实验二

实验一发现可见性缩小不同反应类型间的偏好差异，使偏好更加一致，提示可见性可能让被试更加遵从某种社会规范。基于这一结果，实验二进一步操纵社会规范，探究社会规范对行为可见增加利他偏好的作用。

3.1 方法

3.1.1 被试

收集数据前，使用 G*Power 3.1.9.7 计算预期样本量，基于 2×2 重复测量方差分析的 F 检验，当效力为 0.8、显著性水平为 0.05、效应量为 0.25 时，所需样本量为 34。实际研究中，共招募 54 名成年大学生被试(男性 23 人)，其中 1 人因反应方式刻板(所有题目均选择左侧选项)而被排除，剩余有效被试 53 人(男性 22 人)，年龄在 18~26 岁之间，$M = 20.33$，SD

= 1.82。所有被试母语均为汉语，右利手，视力或矫正视力正常。所有被试均在实验前填写知情同意书，知晓实验的潜在收益与风险，自愿参与实验，并在实验后获得金钱报酬。

3.1.2 实验设计和流程

实验为 2(行为可见性：可见 vs. 不可见) × 2(社会规范：利他 vs. 非利他)的被试内设计。自变量为可见性和社会规范，因变量为被试的利他程度。实验一验证了行为可见增加利他偏好的作用在选择和评分条件下都成立，因此，在实验二中，仅采用选择条件进行研究。

可见性的操纵与实验一相同。对于社会规范，根据描述性规范的定义(Elster, 1989)，通过“之前参加实验的被试选择不公平选项的比例”操纵利他或非利他的社会规范(见图 5，以每个选项背景的灰条长度和右侧的百分数表示选择比例)。具体而言，分别在 AIA 为横轴、DIA 为纵轴的坐标系的第四象限(偏好利他)和第二象限(偏好非利他)各选择 5 个点作为虚拟的偏好组合(见图 5，利他象限选择的点为[0.5, -0.5,]，[0.6, -0.2]，[0.2, -0.6]，[0.3, -0.4]，[0.4, -0.3]，非利他象限选择的点为[-0.5, 0.5,]，[-0.6, 0.2]，[-0.2, 0.6]，[-0.3, 0.4]，[-0.4, 0.3])，此后按照“2.1.3 测量指标”所述的选择效用模型，用每一个虚拟的偏好组合随机生成 10 套虚拟被试的选择数据(即每个规范条件下各生成 50 个虚拟被试的数据)，然后分别计算不同规范下，对于每一次选择，50 个虚拟被试选择不公平选项的比例。为使社会规范的被试内设计更加合理，实验中，被试被告知他们将依次登入不同的游戏服务器，并将看到来自这个服务器的被试选择每一个选项的比例。其中，服务器 1 对应于利他的社会规范，服务器 2 对应于非利他的社会规范。实验后，他们的选择数据也将被同步到这两个服务器中，并呈现给之后参加实验的被试。这样的设置让被试先后处于不同的社会群体中，并感知到该群体中大多数人的行为方式。

实验二也为区块设计，两个自变量组成 2×2 四个区块，即利他-可见，利他-不可见，非利他-可见，非利他-不可见。相同社会规范的区块连续出现，不同社会规范的呈现顺序在被试间随机。在同一社会规范内部，可见和不可见区块的呈现顺序随机。每一区块开始前，屏幕提示该区块的可见性和服务器类型。

实验二的实验流程和实验一的选择任务大致相同，区别在于实验二引入了社会规范的操纵。为了让被试对当前的社会规范形成稳定印象，每一个区块开始时，先进行 3 轮游戏的练习，再进行 12 轮正式游戏(共 48 次二项选择)，每一轮游戏包含一道加减法计算题和 4 次连续金钱分配任务。实验二具体选项设置见本文的线上公开数据

(https://github.com/psych575/open-data-and-code-for-xb21-575.git)。

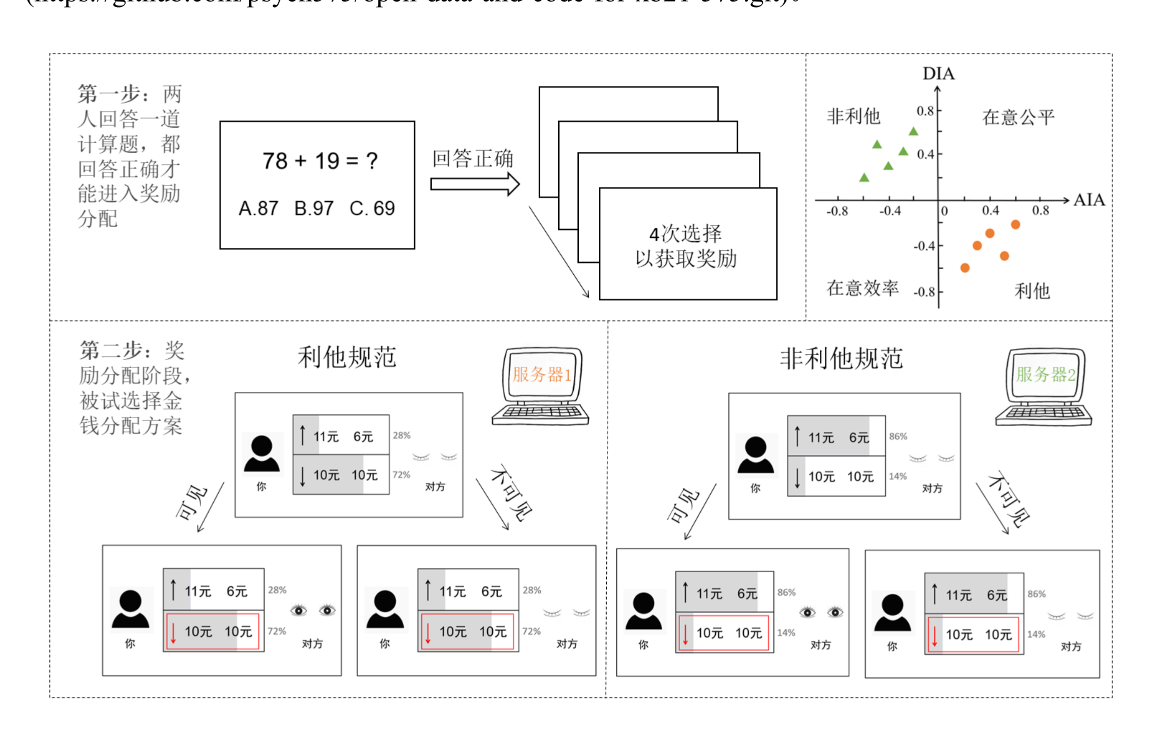


图 5 实验二流程

3.1.3 测量指标

与实验一相同，实验二通过不依赖模型的指标与基于模型的参数测量利他程度。

在不依赖模型的指标中，为了更直观表示利他程度，使用四个条件下被试选中试次中他人的平均收益衡量利他程度，称为平均分给对方的点数。被试平均分给对方的点数越多，说明利他程度越高。此外，与实验一的选择条件相同，分别计算四个条件下，在优势和劣势情况下，选中选项的自我他人收益之差的绝对值的均值(简称自我他人收益之差)。

使用不公平厌恶模型(Fehr & Schmidt, 1999)对被试的行为进行建模。假定被试对不公平选项的效用为：

$$U_{unequal} = M_s - \alpha \cdot max(M_s - M_o,\ 0) - \beta \cdot max(M_o - M_s, 0) \quad (8)$$

分别对利他-可见、利他-不可见、非利他-可见和非利他-不可见四个条件下的 AIA 和 DIA 进行估计，对应的 AIA 依次为α_1，α_2，α_3和α_4，DIA 依次为β_1，β_2，β_3和β_4。效用模型的其他设定和实验一的选择条件相同。此模型称为利他-非利他分离模型。此外，还假设不同社会规范下参数值相同，AIA 和 DIA 仅受到可见性的影响，从而建立利他-非利他联合模型。

使用 R 4.0.2 与 JAGS 4.3.0，建立分层贝叶斯模型(Hierarchical Bayesian Model, HBM)进

行参数拟合。假设不同条件下的 AIA, DIA 以及温度参数分布存在重合，即$\alpha_2=\alpha_1+\triangle\alpha_2$，$\alpha_3=\alpha_1+\triangle\alpha_3$，$\alpha_4=\alpha_1+\triangle\alpha_4$，$\beta_2=\beta_1+\triangle\beta_2$，$\beta_3=\beta_1+\triangle\beta_3$，$\beta_4=\beta_1+\triangle\beta_4$，$\lambda_2=\lambda_1+\triangle\lambda_2$，$\lambda_3=\lambda_1+\triangle\lambda_3$，$\lambda_4=\lambda_1+\triangle\lambda_4$。对每名被试，同时拟合 12 个参数，即$\alpha_1$、$\triangle\alpha_2$、$\triangle\alpha_3$、$\triangle\alpha_4$、$\beta_1$、$\triangle\beta_2$、$\triangle\beta_3$、$\triangle\beta_4$、$\lambda_1$、$\triangle\lambda_2$、$\triangle\lambda_3$、$\triangle\lambda_4$。分层贝叶斯模型的其他设置和实验一的选择条件相同。

3.2 结果

本研究使用 IBM SPSS Statistics 20，MATLAB® R2018a 和 R 4.0.2 进行数据分析与统计检验，结合不依赖模型的结果和计算建模以验证研究假设。

3.2.1 不依赖模型的结果

首先，以可见性和社会规范为自变量，以平均分给他人的点数为因变量，进行 2×2 重复测量方差分析。结果显示，可见性的主效应显著，$F(1,52)=12.38$，$p=0.001$，$\eta_p^2=0.19$，可见条件下分给他人的点数多于不可见条件；社会规范的主效应显著，$F(1,52)=53.74$，$p<0.001$，$\eta_p^2=0.53$，利他规范下分给他人的点数多于非利他规范；可见性与社会规范的交互作用边缘显著 $F(1,52)=3.09$，$p=0.085$，$\eta_p^2=0.06$。简单主效应分析发现，在利他规范下，可见条件下平均分给他人点数更多($D=0.37$，$p=0.001$)；而在非利他规范下，可见和不可见条件下平均分给他人的点数没有显著差异($D=0.12$，$p=0.185$) (图 6c)。这一结果直接说明可见性对利他程度的影响依赖于利他的社会规范。

其次，以优势条件下自我他人收益之差为因变量，以可见性和社会规范为自变量，进行 2×2 重复测量方差分析。结果显示，可见性的主效应显著，$F(1,52)=32.96$，$p<0.001$，$\eta_p^2=0.39$，可见条件自我他人收益之差小于不可见条件；社会规范的主效应显著，$F(1,52)=16.95$，$p<0.001$，$\eta_p^2=0.25$，利他规范下自我他人收益之差小于非利他规范；可见性与社会规范的交互作用显著，$F(1,52)=5.04$，$p=0.029$，$\eta_p^2=0.09$。简单主效应分析发现，在利他规范下，可见条件下自我他人收益之差显著小于不可见条件($D=-0.68$，$p<0.001$)，说明可见时更不能容忍优势不公平；在非利他规范下，同样可见时更不能容忍优势不公平($D=-0.34$，$p=0.001$) (图 6a)。

以劣势条件下自我他人收益之差(绝对值)为因变量，以可见性和社会规范为自变量进行方差分析。结果发现，可见性的主效应边缘显著，$F(1,52)=3.08$，$p=0.085$，$\eta_p^2=0.06$，可

见条件下自我他人收益之差大于不可见条件；社会规范的主效应显著，$F(1,52) = 42.38$，$p < 0.001$，$\eta_p^2 = 0.45$，利他规范下自我他人收益之差大于非利他规范；可见性与社会规范的交互作用不显著，$F(1,52) = 2.01$，$p = 0.162$，$\eta_p^2 = 0.04$。然而，配对样本 t 检验发现，在利他规范下，可见条件下的自我他人收益之差边缘显著地大于不可见条件($t = 2.06$，$p = 0.088$，Bonferroni 矫正，Cohen's $d = 0.23$)，说明可见时更能容忍劣势不公平；而在非利他规范下，可见和不可见条件下自我他人收益之差没有显著差异($t = 0.44$，$p = 0.663$，Cohen's $d = 0.04$)(图 6b)。

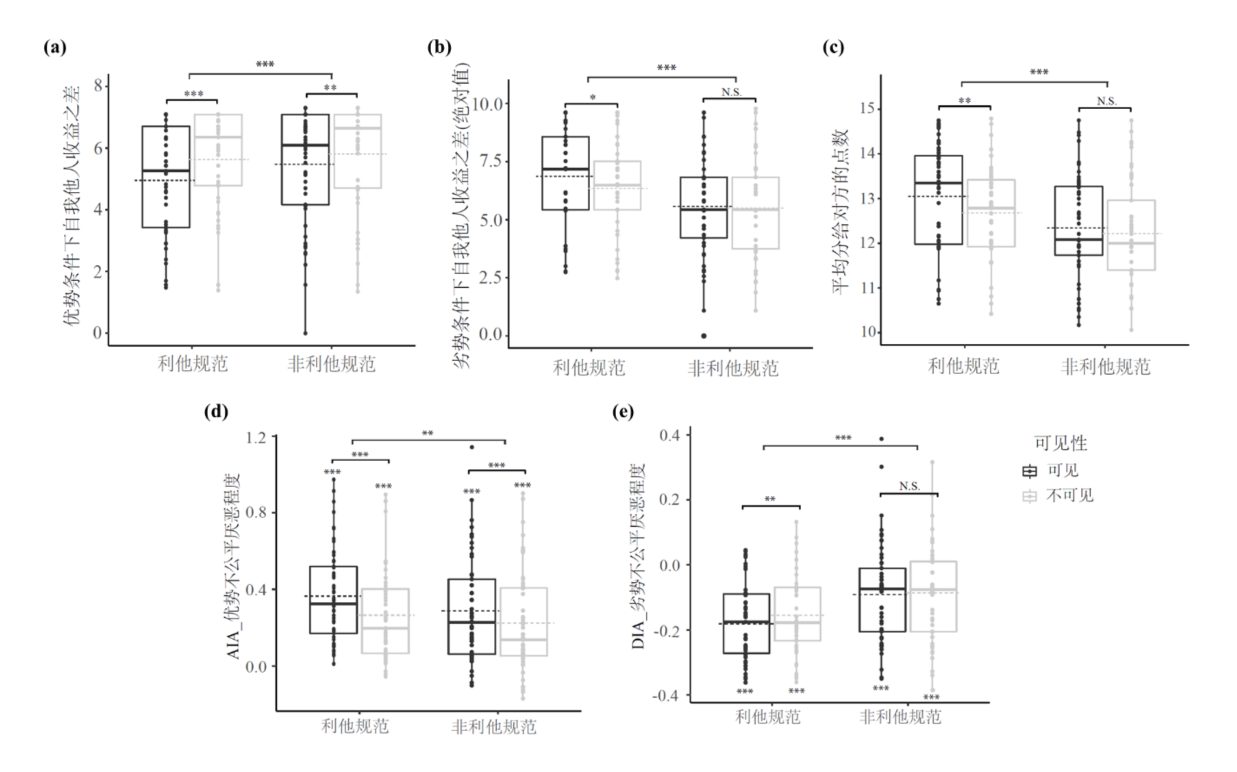


图 6 实验二结果 (a)(b)优势和劣势情况下，自我他人收益之差；**(c)**平均分给他人的点数；**(d)(e)**分层贝叶斯模型拟合的 AIA 和 DIA。图中，箱图内部的水平实线代表中位数，水平虚线代表均值。(d)(e)中，每个箱图上方的显著性表示该条件下参数显著大于零，下方的显著性表示该条件下参数显著小于零。*: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$。

3.2.2 计算建模结果

模型比较显示，利他-非利他分离模型(DIC = 4374.87)优于利他-非利他联合模型(DIC = 4680.11)，因此采用利他-非利他分离模型进行后续分析。利他-非利他分离模型的参数拟合结果见图 6(d 和 e)。

为检验可见性与社会规范对 AIA 和 DIA 的影响，分别以 AIA、DIA 为因变量，以可见

性和社会规范为自变量，进行 2×2 重复测量方差分析。分析主要关注可见性与社会规范的交互作用，以探究可见性对 AIA 和 DIA 的影响是否受到社会规范的调节。

以 AIA 为因变量进行方差分析发现，可见性的主效应显著，$F(1,52) = 86.87$，$p < 0.001$，$\eta_p^2 = 0.63$，可见条件 AIA 大于不可见条件；社会规范的主效应显著，$F(1,52) = 39.47$，$p < 0.001$，$\eta_p^2 = 0.43$，利他规范下 AIA 大于非利他规范；可见性与社会规范的交互作用边缘，$F(1,52) = 3.82$，$p = 0.056$，$\eta_p^2 = 0.068$。简单主效应分析发现，利他规范下，可见条件 AIA 大于不可见条件($D = 0.10$，$p < 0.001$)；非利他规范下，可见条件 AIA 也大于不可见条件($D = 0.063$，$p < 0.001$)。

以 DIA 为因变量进行方差分析发现，可见性的主效应显著，$F(1,52) = 4.82$，$p = 0.033$，$\eta_p^2 = 0.085$，可见条件 DIA 小于不可见条件；社会规范的主效应显著，$F(1,52) = 59.96$，$p < 0.001$，$\eta_p^2 = 0.54$，利他规范下 DIA 小于非利他规范；可见性与社会规范的交互作用不显著，$F(1,52) = 2.55$，$p = 0.116$，$\eta_p^2 = 0.047$。然而，配对样本 t 检验发现，在利他规范下，可见条件下的 DIA 显著小于不可见条件($t = -2.97$，$p = 0.008$，Bonferroni 矫正，Cohen's $d = 0.26$)，说明可见时更能容忍劣势不公平；而在非利他规范下，可见和不可见条件下的 DIA 没有显著差异($t = -0.51$，$p = 0.612$，Cohen's $d = 0.034$)。

以上对计算建模估计参数的方差分析结果和不依赖模型的结果相呼应：在利他规范下，相比于不可见条件，在可见条件下 AIA 上升，DIA 下降；而在非利他规范下，可见性对 AIA 和 DIA 的影响减弱，说明行为可见增加利他偏好的效应依赖于利他的社会规范。

4 讨论

本研究使用行为实验构建社会决策情景，探究在二人分配的修改版独裁者游戏中，接受者对分配行为的可见性对分配者利他偏好的影响。研究结合计算建模，使用不公平厌恶模型量化社会偏好，并结合优势与劣势不公平厌恶的变化来揭示不同情境下利他偏好的变化。实验一采用选择和评分两种反应类型测量社会偏好，发现无论选择还是评分，相比于行为不可见，行为可见时被试的利他程度上升，说明行为可见增加利他偏好，且该效应在不同反应类型间具有一致性；此外，相比于评分，在选择时被试更加在意分配的效率。实验二探究了特定社会规范对行为可见增加利他偏好的影响，发现社会规范和行为可见性对利他偏好的影响存在交互作用：相比于非利他规范，在利他规范下，行为可见增加利他偏好的程度更大。

本研究的发现为利他行为的信号功能假说提供了实验证据支持。本研究基于竞争利他假说和高成本信号理论，即认为人们利他的目的在于传递自己亲社会特质的信号，进而建立良好的社会形象(Ginits et al., 2001; Hardy & Van Vugt, 2006)。Hardy 和 Van Vugt(2006)的研究仅验证了行为可见对于人们在公共物品游戏中的利他行为的影响。本研究将可见性的操纵拓展到二人独裁者游戏中，发现当分配行为能够被接受者看到时，分配者更少选择优势选项，更多选择劣势选项，总体表现出更多的利他行为。反之，当行为不可见时，人们的利他行为减少，甚至出现自利倾向：在实验一的选择条件，不可见时 AIA 值为负，提示此时人们倾向优势不公平。这一结果为利他的策略性动机提供支持(Böckler et al., 2016)，并暗示当行为不可见时，人们会策略性地减少利他行为，增加自我收益。本研究将社会心理学领域解释利他的理论和社会决策领域的利他偏好相结合，为解释利他的功能与动机提供思路。

同时，本研究弥补了关于“可见性(或被观察)”研究的空缺。已有较多研究关注“感知到被观察”、“身份对他人可知”和“第三方观察”等方面 (Bradley et al., 2018)，但少有研究关注“行为可见”和“行为被接受方观察”的决策情境。少数研究曾探讨类似话题，但其对可见性的操纵较为不直观(Andreoni & Bernheim, 2009)。本研究使用人际互动情境操纵“行为是否被接受方可见”，实验一和实验二均证明行为可见促进利他偏好，为验证该结论提供充分证据。

需要注意的是，本研究未定量地操纵和探讨“行为对接受者可见”与其他观察类型研究的具体差异。对于观察的程度深浅，Bradley 等人(2018)的研究发现，在“感知到被观察”“行为被观察”和“行为及身份均被观察”这三类研究中，可观察性对亲社会行为的总体影响大小没有显著差别。然而，不同的观察类型涉及到的心理机制可能不同。如，使用眼睛图片诱发的被观察感受到唤起程度与刺激呈现时长等底层感知觉因素的影响(Hesslinger et al., 2017)；而行为被观察时，人们对“我的行为被他人看见”的事实进行加工，可能涉及较多社会认知过程；身份可见的互动更多发生在熟人之间，可能涉及更多互惠。因此，仍有必要在研究中区分不同的观察程度，并谨慎解释其机制。对于观察者的身份，Bradley 等人(2018)的研究对比了“利益无关的第三方观察”和“其他实验参与者观察”这两种类型，发现第三方观察比其他参与者观察更大程度地促进分配者的利他行为；然而，实验参与者可能担任接受者的身份，进而涉及与分配利益相关、更加在意分配结果或有更多互惠的可能性，应该比第三方观察带来更大程度的利他促进效应。这样的差异可能源于 Bradley 等人(2018)的研究中纳入的“其他实验参与者观察”的研究并非是严格控制的关于“接受者可见”的研究，而本研究填补了这一空缺，为后续研究中探索不同观察者身份的差异提供更多实证证据。

此外，实验一还发现不同反应类型对社会偏好的影响。相比于评分条件，在选择条件下个体 AIA 和 DIA 均下降，总体表现出更加在意分配的效率。Choshen-Hillel 和 Yaniv(2011; 2012)认为，在选择时由于个体有权力做出决策、能够实际影响自己和他人收益，会更在意分配的效率，即最大化总收益；而在评分时，个体更在意分配的公平性，对自我他人收益之差较小的情况更加满意。本研究结果与此观点基本一致。

实验一还发现行为可见性与反应类型存在交互作用，在可见条件下，选择与评分之间的社会偏好差异减小，说明行为可见能够让个体在各个情境下的偏好与策略更趋一致。这一结果与 Nettle 等人(2013)的观点一致，即被观察感减少社会行为的变异性。我们推断这种行为可见减少行为变异性的现象，很可能反映了可见情境让人们更加遵循某种规范的作用。在人类社会中，利他便是一个被广为接受的社会规范(Pereda et al., 2017)，因此当行为对他人可见时，人们可能更加遵循利他的社会规范，进而做出更多利他行为。

为验证该假设，实验二操纵社会规范，施加利他或非利他的社会规范。结果发现，当存在利他规范时，行为可见显著增加利他偏好；而当存在非利他规范时，这一效应减小或消失。这一结果说明行为可见增加利他偏好的作用依赖于利他社会规范的存在，为解释行为可见增加利他偏好的机制提供了思路。同时，本研究发现社会规范的主效应较为明显，说明社会规范对社会偏好影响较大，再次验证了社会规范和社会偏好的紧密关系(Kimbrough & Vostroknutov, 2016)。相比之下，一些关于“感知到被观察”的研究也试图探究社会规范对于解释被观察感增加利他行为的作用(Fathi et al., 2014; Kawamura & Kusumi, 2017; Oda & Ichihashi, 2016)，但并未得到确定性的结论。这些研究大多采用募捐的方式，并在募捐箱上粘贴眼睛图片以创造被观察感，并通过改变募捐箱中已有的金钱数量来操纵社会规范。结果发现，相比于募捐箱中金钱面值较小(即存在非利他规范)，当募捐箱中金钱面值较大时(即存在利他规范)，眼睛图片带来的募捐数额增加并未明显更大(Fathi et al., 2014; Oda & Ichihashi, 2016)；而另一些研究即使发现了社会规范的影响，但结果未得到重复(Kawamura & Kusumi, 2017)。以上研究说明感知到被观察增加利他行为的效应难以被社会规范解释。相较而言，本研究发现社会规范对于行为可见增加利他偏好的效应具有显著的调节作用，说明行为可见时被试更加遵从利他社会规范从而做出更多利他行为。这一差别可能源于“行为为接收者所见”与“感知到被观察”的心理过程差异：感知到被观察和唤起程度等因素有关(Hesslinger et al., 2017)，可能较少受到社会规范等社会因素的影响；而行为可见时人们对社会情境可能进行了更多认知加工，进而更能将社会规范的信息纳入决策过程。

本研究也存在一些不足与改进空间。本研究通过操纵社会规范，发现社会规范对行为可

见增加利他偏好的调节作用，但我们并不排除存在其他机制的可能，如声誉和共情。一些研究也提出声誉对于促进利他等亲社会行为的重要作用(Hardy & Van Vugt, 2006; Roberts, 1998; Suzuki & Akiyama, 2005; Sylwester & Roberts, 2010; Van Vugt et al., 2012; Wedekind & Braithwaite, 2002)，在行为可见时，人们通过利他行为来为自己积累声誉，进而构建良好的社会形象(Milinski et al., 2002)。相比之下，行为不可见时，人们即使做出利他行为，也无法提升声誉，因此行为可见可能通过对声誉的追求进而促进利他行为。对比声誉和遵从社会规范这两种因素，虽然从实验操纵和结果解释上可能有所不同(e.g., Piazza & Bering, 2008)，但在一定程度上都是社会赞许性的体现，即人们倾向于做出符合社会期望的行为。这其实也体现了利他的信号功能，即决策者通过利他行为向社会传递自己的亲社会性。此外，本研究中接受者是否可见的影响，还可能与共情以及心理理论有关(Brüne & Brüne-Cohrs, 2006; Elliott et al., 2011)。当接受者能够看到决策者的行为时，决策者可能更能意识到接受者的处境，体会其情绪并换位思考，进而做出更多利他行为。已有研究表明共情和心理理论都与社会偏好有关(e.g., Tsoi & McAuliffe, 2020; Wu & Han, 2021)，但没有研究探究过这二者在行为可见增加利他偏好中的作用，后续研究可进行进一步探索。

此外，本研究使用选择和评分两种方式测量社会偏好，并假设选择和评分的行为都基于相同的内在效用函数形式，进而对个体的社会偏好进行计算建模，比较不同反应类型下的模型参数。我们的比较建立在目前被广泛使用的社会效用函数模型上，对结果的解释基于实验设计的具体形式和估计的参数结果，后续研究可以进一步探索其他适合的量化指标与比较方式。

参考文献

Andreoni, J., & Bernheim, B. D. (2009). Social image and the 50-50 norm: A theoretical and experimental analysis of audience effects. *Econometrica, 77*(5), 1607–1636.

Andreoni, J., & Petrie, R. (2004). Public goods experiments without confidentiality: A glimpse into fund-raising. *Journal of Public Economics, 88*(7), 1605–1623.

Agerström, J., Carlsson, R., Nicklasson, L., & Guntell, L. (2016). Using descriptive social norms to increase charitable giving: The power of local norms. *Journal of Economic Psychology, 52*, 147–153.

Bateson, M., Nettle, D., & Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters, 2*(3), 412–414.

Böckler, A., Tusche, A., & Singer, T. (2016). The structure of human prosociality: Differentiating altruistically motivated, norm motivated, strategically motivated, and self-reported prosocial behavior. *Social Psychological*

and Personality Science, 7(6), 530–541.

Bradley, A., Lawrence, C., & Ferguson, E. (2018). Does observability affect prosociality? *Proceedings of the Royal Society. B, Biological Sciences, 285*(1875), 20180116.

Brüne, M., & Brüne-Cohrs, U. (2006). Theory of mind—evolution, ontogeny, brain mechanisms and psychopathology. *Neuroscience and Biobehavioral Reviews, 30*(4), 437–455.

Burnham, T. C., & Hare, B. (2007). Engineering human cooperation: Does involuntary neural activation increase public goods contributions? *Human Nature, 18*(2), 88–108.

Choshen-Hillel, S., & Yaniv, I. (2011). Agency and the construction of social preference: Between inequality aversion and prosocial behavior. *Journal of Personality and Social Psychology, 101*(6), 1253–1261.

Choshen-Hillel, S., & Yaniv, I. (2012). Social preferences shaped by conflicting motives: When enhancing social welfare creates unfavorable comparisons for the self. *Judgment and Decision Making, 7*(5), 618–627.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology, 58*(6), 1015–1026.

Dana, J., Cain, D. M., & Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes, 100*(2), 193–201.

Dempsey, R. C., McAlaney, J., & Bewick, B. M. (2018). A critical appraisal of the social norms approach as an interventional strategy for health-related behavior and attitude change. *Frontiers in Psychology, 9*, 2180.

Elliott, R., Bohart, A. C., Watson, J. C., & Greenberg, L. S. (2011). empathy. *Psychotherapy, 48*(1), 43–49.

Elster, J. (1989). Social norms and economic theory. *The Journal of Economic Perspectives, 3*(4), 99–117.

Fathi, M., Bateson, M., & Nettle, D. (2014). Effects of watching eyes and norm cues on charitable giving in a surreptitious behavioral experiment. *Evolutionary Psychology, 12*(5), 878–887.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior, 25*(2), 63–87.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics, 114*(3), 817–868.

Fox, J., & Guyer, M. (1978). "Public" choice and cooperation in n-person prisoner's dilemma. *The Journal of Conflict Resolution, 22*(3), 469–481.

Gao, X., Yu, H., Sáez, I., Blue, P. R., Zhu, L., Hsu, M., & Zhou, X. (2018). Distinguishing neural correlates of context-dependent advantageous- and disadvantageous-inequity aversion. *Proceedings of the National Academy of Sciences, 115*(33), E7680–E7689.

Gintis, H., Smith, E. A., & Bowles, S. (2001). Costly signaling and cooperation. *Journal of Theoretical Biology, 213*(1), 103–119.

Glimcher, P. W., & Fehr, E. (2013). *Neuroeconomics: Decision making and the brain: Second edition.* Academic Press.

Haley, K. J., & Fessler, D. M. T. (2005). Nobody's watching? Subtle cues affect generosity an anonymous economic game. *Evolution and Human Behavior, 26*(3), 245–256.

Hamilton, A., & Lind, F. (2016). Audience effects: What can they tell us about social neuroscience, theory of mind

and autism? *Culture and Brain, 4*(2), 159–177.

Hardy, C. L., & Van Vugt, M. (2006). Nice guys finish first: The competitive altruism hypothesis. *Personality and Social Psychology Bulletin, 32*(10), 1402–1413.

Hesslinger, V. M., Carbon, C., & Hecht, H. (2017). The sense of being watched is modulated by arousal and duration of the perceptual episode. *i-Perception, 8*(6), 2041669517742179.

Hu, Y., He, L., Zhang, L., Wölk, T., Dreher, J., & Weber, B. (2018). Spreading inequality: Neural computations underlying paying-it-forward reciprocity. *Social Cognitive and Affective Neuroscience, 13*(6), 578–589.

Jerdee, T. H., & Rosen, B. (1974). Effects of opportunity to communicate and visibility of individual decisions on behavior in the common interest. *Journal of Applied Psychology, 59*(6), 712–716.

Kawamura, Y., & Kusumi, T. (2017). The norm-dependent effect of watching eyes on donation. *Evolution and Human Behavior, 38*(5), 659–666.

Kimbrough, E. O., & Vostroknutov, A. (2016). norms make preferences social. *Journal of the European Economic Association, 14*(3), 608–638.

Lacetera, N., & Macis, M. (2010). Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme. *Journal of Economic Behavior & Organization, 76*(2), 225–237.

Li, O., Xu, F., & Wang, L. (2018). Advantageous inequity aversion does not always exist: The role of determining allocations modulates preferences for advantageous inequity. *Frontiers in Psychology, 9*, 749–749.

Liu, Y., Li, S., Lin, W., Li, W., Yan, X., Wang, X., Pan, X., Rutledge, R. B., & Ma, Y. (2019). Oxytocin modulates social value representations in the amygdala. *Nature Neuroscience, 22*(4), 633–641.

Loewenstein, G. F., Thompson, L., & Bazerman, M. H. (1989). Social utility and decision making in interpersonal contexts. *Journal of Personality and Social Psychology, 57*(3), 426–441.

McAuliffe, K., Blake, P. R., Kim, G., Wrangham, R. W., & Warneken, F. (2013). Social influences on inequity aversion in children. *PloS One, 8*(12), e80966.

McAuliffe, K., Blake, P. R., & Warneken, F. (2020). Costly fairness in children is influenced by who is watching. *Developmental Psychology, 56*(4), 773–782.

McAuliffe, K., Jordan, J. J., & Warneken, F. (2015). Costly third-party punishment in young children. *Cognition, 134*, 1–10.

McBride, M., & Ridinger, G. (2021). Beliefs also make social-norm preferences social. *Journal of Economic Behavior & Organization, 191*(3), 765–784.

Milinski, M., Semmann, D., & Krambeck, H. (2002). Reputation helps solve the 'tragedy of the commons'. *Nature, 415*(6870), 424–426.

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. J. (2011). Measuring social value orientation. *Judgment and Decision Making, 6*(8), 771–781.

Nettle, D., Harper, Z., Kidson, A., Stone, R., Penton-Voak, I. S., & Bateson, M. (2013). The watching eyes effect in the dictator game: It's not how much you give, it's being seen to give something. *Evolution and Human Behavior, 34*(1), 35–40.

Oda, R., & Ichihashi, R. (2016). Effects of eye images and norm cues on charitable donation: A field experiment in

an izakaya. *Evolutionary Psychology, 14*(4), 147470491666887.

Oda, R., Niwa, Y., Honma, A., & Hiraishi, K. (2011). An eye-like painting enhances the expectation of a good reputation. *Evolution and Human Behavior, 32*(3), 166–171.

Palacios, J., Fan, Y., Yoeli, E., Wang, J., Chai, Y., Sun, W., Rand, D. G., & Zheng, S. (2022). Encouraging the resumption of economic activity after COVID-19: Evidence from a large scale-field experiment in China. *Proceedings of the National Academy of Sciences, 119*(5), e2100719119.

Pereda, M., Brañas-Garza, P., Rodríguez-Lara, I., & Sánchez, A. (2017). The emergence of altruism as a social norm. *Scientific Reports, 7*(1), 9684.

Pfattheicher, S., Nielsen, Y. A., & Thielmann, I. (2022). Prosocial behavior and altruism: A review of concepts and definitions. *Current Opinion in Psychology, 44*, 124–129.

Piazza, J., & Bering, J. M. (2008). Concerns about reputation via gossip promote generous allocations in an economic game. *Evolution and Human Behavior, 29*(3), 172–178.

Raiffa, H., & Schlaifer, R. (1961). *Applied statistical decision theory*. University College London: Biometrika Office.

Raihani, N. J., & Bshary, R. (2012). A positive effect of flowers rather than eye images in a large-scale, cross-cultural dictator game. *Proceedings of the Royal Society B: Biological Sciences, 279*(1742), 3556–3564.

Raihani, N. J., & McAuliffe, K. (2014). Dictator game giving: The importance of descriptive versus injunctive norms. *PloS One, 9*(12), e113826.

Roberts, G. (1998). Competitive altruism: From reciprocity to the handicap principle. *Proceedings of the Royal Society B: Biological Sciences, 265*(1394), 427–431.

Sáez, I., Zhu, L., Set, E., Kayser, A., & Hsu, M. (2015). Dopamine modulates egalitarian behavior in humans. *Current Biology, 25*(7), 912–919.

Suzuki, S., & Akiyama, E. (2005). Reputation and the evolution of cooperation in sizable groups. *Proceedings of the Royal Society B: Biological Sciences, 272*(1570), 1373–1377.

Sylwester, K., & Roberts, G. (2010). Cooperators benefit through reputation-based partner choice in economic games. *Biology Letters, 6*(5), 659–662.

Tian, Y. (2016). *Research on game experiments of fairness preference: The influence of dominance status, punishment situation and fairness perception on the fairness preference.* [Masteral dissertation, Harbin Engineering University]. Harbin.

[田莹(2016). *公平偏好的博弈实验研究——支配地位、惩罚情境和公平感知对公平偏好的影响.* [硕士学位论文，哈尔滨工程大学]. 哈尔滨.]

Tsoi, L., & McAuliffe, K. (2020). Individual differences in theory of mind predict inequity aversion in children. *Personality and Social Psychology Bulletin, 46*(4), 559–571.

van Lange, P. A. M. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology, 77*(2), 337–349.

van Lange, P. A. M., de Bruin, E. M. N., Otten, W., & Joireman, J. A. (1997). The development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence. *Journal of Personality and*

Social Psychology, 73(4), 733–746.

Van Vugt, M., Roberts, G., & Hardy, C. (2012). Competitive altruism: A theory of reputation-based cooperation in groups. In Louise Barrett, Robin Dunbar (ed.), *Oxford Handbook of Evolutionary Psychology* (pp. 531-540). Oxford University Press.

Wedekind, C., & Braithwaite, V. A. (2002). The long-term benefits of human generosity in indirect reciprocity. *Current Biology, 12*(12), 1012–1015.

West, S. A., Griffin, A. S., & Gardner, A. (2007). Social semantics: Altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology, 20*(2), 415–432.

Wu, T., & Han, S. (2021). Neural mechanisms of modulations of empathy and altruism by beliefs of others' pain. *eLife, 10*, e66043.

Social norm modulates the enhancement effect of behavioral visibility on altruistic preference

HUANG Xinru[1]; LI Jian[1, 2]; NI Yinmei[1]

([1] *School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, 100871, China)*

([2] *IDG/McGovern Institute for Brain Research, Peking University, Beijing* 100871, *China*)

Abstract

Background: In social economic decisions, people not only care about their own payoffs but also the payoffs of others, a tendency termed altruistic preference. Numerous studies have shown that the sheer sense of being observed is sufficient to augment subjects' altruistic choices. However, whether subjects' altruistic behavior can be modulated by other stake-holders in the decision context remains unclear. In this study, we provide experimental evidence about the effects of visibility from receivers and social norms on the altruistic preference of the deciders in two studies. First, we confirmed the visibility effect originating from receivers on deciders' altruistic preference in experiment 1. In experiment 2, we further showed that social norms modulated the effect of behavior visibility on deciders' altruistic preference, suggesting a potential avenue via which social norms influences the relationship between behavioral visibility and altruistic preference.

Study 1: Study 1 implemented a two (visibility: visible vs. invisible) x two (reaction type: choice vs. rating) x two (inequity aversion: AIA vs. DIA) within-subject design. We recruited 38 participants and they were required to either choose from two reward allocation options (choice task) with another partner, or rate how satisfied concerning a particular allocation (rating task) in a dictator game (DG). Participants' behavior was either observed by their "partners" (visible condition) or remained private (invisible condition). We provided both model-free and model-based evidence for the effects of visibility on altruistic preference. Compared to the invisible condition, participants exhibited greater altruistic preference when their behavior were visible to the receivers (partners).

This tendency was significant across both choice and rating tasks. In addition, participants cared more about allocation efficiency in the choice task than in the rating task. Finally, visibility alleviated the behavioral discrepancies between the rating and choice tasks, indicating that social preference and choice strategies tend to converge in the visible condition.

Study 2: Study 2 implemented a two (visibility: visible vs. invisible) x two (social norm: altruistic vs. non-altruistic) within-subject design. 53 participants took part in the study with altruistic or non-altruistic social norms. Different social norms were manipulated by presenting the proportions of choosing the unfair options from previous participants: In the altruistic social norm condition, most previous participants chose the option that maximizes others' relative payoffs, while in the non-altruistic condition it was the opposite. Our results showed that in the altruistic social norm condition, visibility significantly increased participants' altruistic preference. However, such effect diminished in the non-altruistic social norm condition.

Conclusion: Our study revealed that deciders' behavioral visibility to receivers increased altruistic preference and promoted altruistic behavior. Furthermore, the altruistic social norm played a modulatory role on the visibility effect, supporting the signaling hypothesis of altruistic preference.

Keywords: social preference, altruism, visibility, reaction type, social norm